

Blind audio source counting and separation of anechoic mixtures using the multichannel complex NMF framework

Sayeh Mirzaei^{a,*}, Hugo Van hamme^a, Yaser Norouzi^b

^a Department of Electrical Engineering, KULeuven, Belgium

^b Department of Electrical Engineering, Amirkabir University, Tehran, Iran

ARTICLE INFO

Article history:

Received 24 November 2014

Received in revised form

28 January 2015

Accepted 8 March 2015

Available online 18 March 2015

Keywords:

Blind Source Separation (BSS)

Complex Non-negative Matrix Factorization (CNMF)

Binary masking

Anechoic mixture

ABSTRACT

In this paper, we address the tasks of audio source counting and separation for a stereo anechoic mixture of audio signals. This will be achieved in two stages. In the first stage, a novel approach is introduced for estimating the number of sources as well as the channel mixing coefficients. For this purpose, a 2-D spectrum is evaluated against both the phase and amplitude differences of the two channels. Hence, obtaining the peak locations of the spectrum yields the number of the sources and the corresponding channel coefficients. In the second stage, an extension of a single channel complex matrix factorization method to multichannel is developed to extract the individual source signals. We find primary estimates of the sources via binary masking and then apply the complex factorization to the complex spectrogram of each source. The obtained factors are then utilized as initial values in the complex multichannel factorization model. We also suggest a method for estimating the number of required components for modeling each source. The separation performance improvement over the conventional methods is investigated by calculating BSS evaluation metrics. The comparison is also carried out in terms of source counting and localization with the recently proposed DeMIX-Anechoic method.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Most audio signals are mixtures of several sources which might be active concurrently. Separation of audio signals is required for several audio processing tasks such as speech recognition, speaker identification, and polyphonic music transcription.

When no prior information of the sources or channel mixing system is available, the task is called Blind Source Separation (BSS). When the mixing coefficients are real value gains, the mixture is termed instantaneous. However in real environments, the mixing process consists of a

linear time-invariant filtering of the source signals as

$$x_i(t) = \sum_{j=1}^J \sum_{\tau=-\infty}^{\infty} a_{ij}(\tau) s_j(t-\tau), \quad i=1,2 \quad (1)$$

where $s_j(t)$, $j=1\dots J$, are the source signals and $a_{ij}(\tau)$ denotes the mixing filter. If the length of the filters a_{ij} are sufficiently smaller than the window length, (1) can be represented in the frequency domain as the following approximation:

$$\mathbf{X}_{ft} = \mathbf{A}(f) \mathbf{S}_{ft} \quad (2)$$

where $\mathbf{X}_{ft} = [X_{1ft} \ X_{2ft}]^T$ indicates the mixture signal complex STFT coefficients on two channels. $\mathbf{A}(f)$ represents the $2 \times J$ matrix containing the channel mixing coefficients and $\mathbf{S}_{ft} = [S_{1ft} \dots S_{Jft}]^T$ denotes the complex spectrogram coefficients of the sources. For the case of anechoic mixtures where no reflection is assumed from the objects

* Corresponding author.

E-mail address: smirzaei@esat.kuleuven.be (S. Mirzaei).

existing in the environment, the complex valued coefficients of matrix \mathbf{A} have frequency independent magnitudes while their phase varies with frequency depending on the microphone array arrangement and the source location w.r.t. the array. The Fourier transform of the mixing filter coefficients can be parameterized as [12]

$$a_{ij}(f) = \kappa_{ij} e^{-i2\pi f \tau_{ij}} \\ \kappa_{ij} = \frac{1}{\sqrt{4\pi r_{ij}}} \quad \text{and} \quad \tau_{ij} = \frac{r_{ij}}{C_s} \quad (3)$$

where r_{ij} specifies the distance between j th source and i th microphone and C_s is the sound propagation velocity.

In a more realistic environment, there are several paths besides the direct path from which the signals impinge on the microphones. The mixtures of this scenario are called convolutive and the \mathbf{A} coefficients cannot be assumed frequency independent anymore in this case.

In this paper, we propose a framework for anechoic mixtures but we also apply it to the mixtures in moderate reverberant environments. We confirm the effectiveness of the model even in the presence of multipath propagation.

Many methods have been developed for blind source separation. They either work based on a known channel mixing system or they try to estimate the mixing system in advance. Most of the channel mixing estimation techniques are based on time difference of arrival (TDOA) estimation in every time–frequency (TF) cell of the STFT representation of the mixture which is followed by clustering of the obtained estimates [1]. In Yilmaz and Rickard [32], Winter et al. [31], and Saab et al. [22], a clustering algorithm is used for finding the clusters around the actual mixing vectors. Besides some drawbacks associated to the spatial aliasing issue, most of these approaches suffer from an important deficiency which is the need that the number of clusters or sources be known in advance. Moreover, the assumption of strong sparsity is an essential requirement of most of these methods. It means that there should be only one dominant active source in each TF cell of the mixture STFT representation. However, the sparsity can be violated either in the presence of noise or when the sources overlap.

Due to the above mentioned disadvantages, we do not opt for clustering based methods. Instead, we obtain a 2-D spectrum against the amplitude and phase differences of the two channels as described in Section 2. The peak locations of this spectrum correspond to the mixing coefficients. This way, we have relaxed the above mentioned sparsity assumption in the sense that we require a sufficient number of sparse TF cells for acquiring clear spectrum peaks corresponding to each source.

Independent Component Analysis (ICA) is a well-known BSS approach. Generally, ICA assumes that source signals are independent and it involves the so-called permutation problem since ICA is applied to separate frequency cells of the STFT representation. Hence, the source orders should be determined per frequency. This issue has been considered and solved in [13,23,19,25]. But ICA is just applicable to over-determined problems where the number of sensors is at least as large as the number of the sources. Under-determined mixtures are more often

separated using TF masking techniques [32,21] or classical sparse approaches such as l_1 -norm [6] or l_p -norm [27] minimization.

Recently, Non-negative Matrix Factorization (NMF) has been applied extensively to several source separation scenarios in a single channel [26,14,30,8,5] and multi-channel [20,10,11,18] setting. NMF or Non-negative Tensor Factorization (NTF) is generally applied to the magnitude spectrogram domains. One of the shortcomings of these structures is that they disregard the phase information of the signal in the sense that they try to approximate the power or magnitude spectrogram of the mixture signal as a product of two non-negative matrices containing spectral components and time activations. Moreover, the number of the sources as well as components modeling each source should be known in advance.

In Kameoka et al. [15], a complex NMF model is introduced defining a mixing model in the complex-spectrogram domain. This model is able to provide a sparse representation of acoustic signals in a single-channel scenario. Utilizing this framework for the source separation task, we still need to know the model order in terms of the number of the sources and the components required for modeling each source.

In our present work, we have extended the method in Kameoka et al. [15] to the multichannel case. Our method differs from that of Sawada et al. [24,4] in the sense that they have proposed techniques for source separation in reverberant environment based on modeling the spatial covariance matrix. In Arberet et al. [4], the contribution of each source to the mixture channels in the time–frequency domain is modeled by a zero-mean Gaussian random vector with a full rank covariance matrix composed of two terms: a variance which represents the spectral properties of the source and which is modeled by a non-negative matrix factorization model and another full rank covariance matrix which encodes the spatial properties of the source contribution in the mixture. The extended complex NMF model in Sawada et al. [24] has been developed to factorize the covariance matrix and the spatial property is associated with each NMF basis. However our proposed method is directly applied to the complex spectrogram of the mixture signal, thus it involves less computation. The estimates of the number of the sources and channel mixing given by the previous stage are exploited as known parameters in our developed model. A primary estimate of each source complex spectrogram is obtained through binary masking. Then, the complex NMF approach proposed in Kameoka et al. [15] is applied for decomposing these spectrograms to the non-negative spectral components and time activation matrices along with a tensor containing the phase information. The obtained factors are then utilized as initial values of the parameters in our proposed multichannel complex NMF approach. We have also developed a scheme for estimating the number of components required for modeling each source. This is achieved through log-likelihood evaluation against different model order values and taking the knee point which can be found by using Bayesian Information Criterion (BIC) metric as the optimal order value.

The novel aspects of our proposed approach can be summarized as follows:

- We have developed a new scheme for estimating the channel mixing coefficients in the anechoic case based on a 2-D spectrum. Applying this approach, both the phase and attenuation differences between the responses of two channels are inferred for each source from the mixture complex spectrograms. Moreover, the number of the peaks emerging in the spectrum provides an estimate of the number of sources.
- The complex NMF model is extended and reformulated to the multichannel case for extracting the individual source signals.
- The optimal model order is chosen based on log-likelihood evaluation against different numbers of model components and applying the BIC metric as the order selection criterion.

The structure of the rest of the paper is as follows: We introduce our proposed spectrum based approach for obtaining the channel mixing coefficients in Section 2. We then explain our developed complex mixture model and the source separation algorithm in Section 3. Various experimental settings and the performance results are presented in Section 4. Finally, we conclude and discuss future research directions in Section 5.

2. Estimating the number of sources and channel mixing coefficients

Several algorithms have already been proposed for channel mixing system estimation for different scenarios of instantaneous, anechoic or convolutive mixtures. The common assumption of most of them is the sparsity of the audio source signals. It means that the source representations in the TF domain do not overlap. This assumption is not necessarily valid for all mixtures. Most of the mentioned methods find the channel estimates through clustering the coefficients obtained for each TF cell. As stated in Section 1, our proposed method is not clustering based and enables us to infer the number of sources as well. In the anechoic case, not only the phase difference between the signals received from two channels should be taken into account, but also the amplitude difference due to the different attenuation of the paths should be considered in the model. While most of the far-field models disregard the difference in the received signal amplitude and just consider the phase difference, we develop a 2-D spectrum for estimating both the phase and amplitude of the channel coefficients. This would be essential especially in the case where the source distances to individual microphones differs considerably.

Our goal is finding a 2-D spectrum against (R_g, θ) from which we can obtain the channel coefficients (magnitude and phase) corresponding to each source. Toward this goal, the subsequent procedures followed. First, τ_g is defined as

$$\tau_g = \frac{d_m \sin(\theta)}{C_s} \quad (4)$$

where d_m is the distance between the microphones. θ denotes the angle of arrival of the signal impinging on the array w.r.t. the array broadside which is assumed uniformly aligned in the interval $[-\pi/2 \ \pi/2]$. In fact, τ_g determines the TDOA between the received signals of the two channels. R_g is used for inferring the ratio of the absolute values of the channel coefficients. We assign uniformly aligned values between $[0 \ 1]$ to R_g . Subsequently, a complex valued function A_g can be defined as below:

$$A_g(R_g, \tau_g, f) = R_g \exp(i2\pi\tau_g f) \quad (5)$$

We can also define a corresponding metric A_{21} as

$$A_{21}(f, t) = \cos \left(\text{atan} \left(\frac{|X_{2ft}|}{|X_{1ft}|} \right) \right) \cdot \exp(i(\angle X_{2ft} - \angle X_{1ft})) \quad (6)$$

which specifies the phase difference as well as the ratio of the magnitudes of the two channels. A_{21} is defined to be located inside the unit circle such that it is comparable with the A_g parameter. X_{ift} , $i = 1, 2$ denotes the complex valued STFT coefficients of channel i in the frequency bin f and time frame t .

The following cost function is then evaluated in each TF cell to give us an estimate of the actual amplitude and phase characteristics of the mixing system:

$$M(R_g, \theta, f, t) = \frac{|A_g - A_{21}|}{2} \quad (7)$$

For increasing the resolution, a monotonically decreasing non-linear function in the range $[0, 1]$ is applied to the M metric inspired from Loesch and Yang [17] which leads to sharper peaks in the derived spectrum:

$$P(R_g, \theta, f, t) = 1 - \tanh(\alpha M(R_g, \theta, f, t)) \quad (8)$$

where α indicates the non-linearity parameter. The 2-D spectrum is then derived by taking summation over all frequency bins and maximization over all time frames:

$$\Gamma(R_g, \theta) = \max_t \sum_f P(R_g, \theta, f, t) \quad (9)$$

Consequently, the phase and amplitude difference measures of two channels can be obtained through finding the peak locations of this 2-D spectrum. Two constraints are applied to the peak finder algorithm. First, the minimum peak distances in terms of angle of arrival (θ) should not exceed 5° . The second constraint is imposed on minimum peak height which should not be smaller than half of the maximum peak value in the spectrum. This way, we have also achieved an estimate of the number of sources in the mixture signal according to their diverse positions which lead to separate peaks emerging in the spectrum. Notice that the peak heights are not proportional to the source signal strengths but to the number of time–frequency cells in which a source is dominant. This can be seen from (9): the derived 2-D spectrum is computed as the sum over all frequencies of values between 0 and 1, measuring agreement of data and hypothesis, and this at the time t where the source can produce the best match. This motivates our choice of threshold for the minimal peak height; implying sources should at some point in time be dominant in at least half the number of cells compared to the most dominant

source. The minimal angular distance between peaks is set based on the practical consideration that sources are spatially distributed.

The inferred peak values of the 2-D spectrum will provide us with the corresponding θ^j and R_g^j associated to the j th source. Subsequently, we can express the estimated channel mixing coefficients with the following form taking the first channel as the reference:

$$\mathbf{A}_{\text{est}}^j(f) = [1 \quad \tan(\text{acos}(R_g^j)) \cdot \exp(i2\pi\tau_g^j f)]^T, \quad j = 1 \dots J, \\ \tau_g^j = \frac{d_m}{C_s} \sin(\theta^j) \quad (10)$$

where J indicates the total estimated number of sources.

A sufficient degree of sparseness is required for our approach to lead to the sharp peaks corresponding to the true channel coefficients. This means that there should be a sufficient number of TF cells with one dominant source in the mixture STFT.

3. Source separation approach

Primary estimates of the individual source spectrograms are obtained through binary masking as explained in Section 3.1. Then, the factors of this primary source spectrograms are obtained using the complex NMF scheme proposed in Kameoka et al. [15]. These factors are then exploited as initial values in our proposed multichannel complex framework to extract the source signals. This framework is described in Section 3.2.

3.1. Primary source estimates and complex factorization

For constructing binary masks, the estimated channel characteristics from Section 2 are exploited in this stage. The active source in each TF cell is decided based on the inferred channel coefficients in the previous stage. The mixture STFT coefficients \mathbf{X}_{ft} are projected onto the subspace spanned by each mixing vector $\mathbf{A}_{\text{est}}^j(f)$ and the source j^* whose projection has largest l_2 -norm is recognized as the active source. The contribution of this source in the given TF cell is taken equal to the channel 1 coefficient, $S_{jft}^{bm} = X_{1ft}$ and that of the other sources is set to zero.

Then, the framework proposed in Kameoka et al. [15] is applied to each source complex spectrogram to extract the recurrent patterns of magnitude spectra and the phase estimates of the constituent components. In contrast to

Kameoka et al. [15], the number of components for modeling each source is not taken predefined and is inferred through applying the BIC order selection metric to log-likelihood values calculated against a different number of components. This is discussed more explicitly in Section 4. Furthermore, we do not assume a sparse prior for time activations in the model because it might not fit to the actual behavior of the parameters.

The following generative model is assumed for the estimated source spectrogram coefficients S_{jft}^{bm} :

$$S_{jft}^{bm} = \sum_{k \in K_j} W_{fk}^{bm} H_{kt}^{bm} e^{i\phi_{kft}^{bm}} + n_{ft} \quad (11)$$

where K_j is the set of indices corresponding to the components used for modeling the j th source. In this model, the columns of the \mathbf{W}^{bm} matrix denote the magnitude spectral components and the rows of the \mathbf{H}^{bm} matrix represent the time-varying activation coefficients of these components. Φ^{bm} accounts for the time-varying phase spectrum of the components. n_{ft} is assumed to be complex Gaussian white noise with mean 0 and variance σ^2 . The update algorithm presented in Kameoka et al. [15] can be used for obtaining the set of unknown parameters $\mathbf{Z}^{bm} = \{\mathbf{W}^{bm}, \mathbf{H}^{bm}, \Phi^{bm}\}$ through maximizing the likelihood function without imposing sparsity to the \mathbf{H}^{bm} elements.

The total number of components, $|K_j|$, required for modeling source j can be inferred via evaluating the likelihood against different order values and choosing the optimal model order using the BIC metric. The factors obtained from binary masked source estimates are utilized in the next stage as initial values of the parameters.

3.2. Complex multichannel matrix factorization framework

We propose to extend the above complex factorization model to the stereo case as follows. The mixture signal is assumed to follow a similar generative model:

$$\mathbf{X}_{ft} = \sum_{k=1}^K W_{fk} H_{kt} e^{i\phi_{kft}} \mathbf{A}_{\text{est}}^{p_k} + \mathbf{n}_{ft} \\ p_k = j \quad \forall k \in K_j, \quad j = 1 \dots J \quad (12)$$

where K represents the total number of components ($K = \sum_j |K_j|$). For later use, we define the vectors \mathbf{Y}_{ft} such that $\mathbf{Y}_{ft} = \sum_{k=1}^K W_{fk} H_{kt} e^{i\phi_{kft}} \mathbf{A}_{\text{est}}^{p_k}$. The idea behind this framework is that the same model components be assumed for both channels. The estimated channel coefficients account for the phase and attenuation difference between channels for individual sources. \mathbf{n}_{ft} is the vector representing the reconstruction error on two channels and is

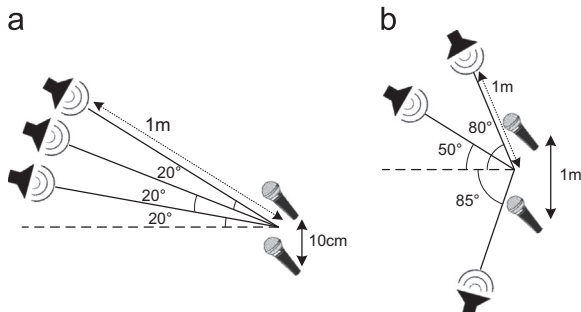


Fig. 1. Source-mic configuration of (a) scenarios 1, 3 and 4, and (b) scenario 2.

Table 1

Experimental parameter setting.

Number of R_g segments	50
Number of θ segments	180
Nonlinearity parameter α	20
Signal duration	10 s
Sampling rate	16 kHz
STFT frame size	1024
STFT frame shift	512
Number of iterations	500

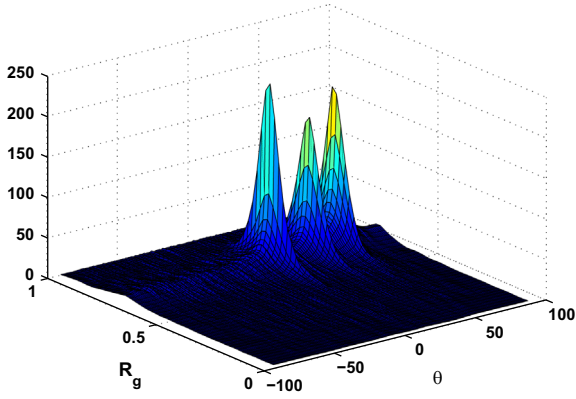


Fig. 2. 2-D spectrum for anechoic mixture of scenario 1.

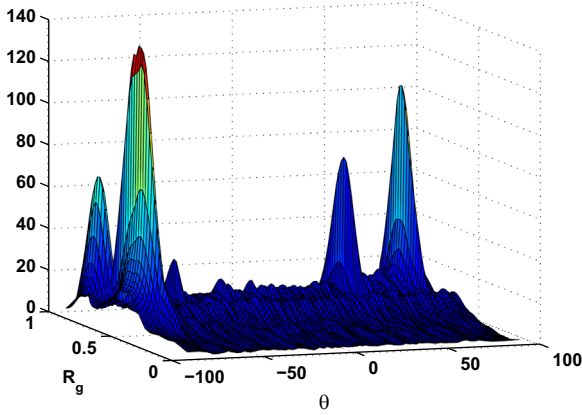


Fig. 3. 2-D spectrum for anechoic mixture of scenario 2.

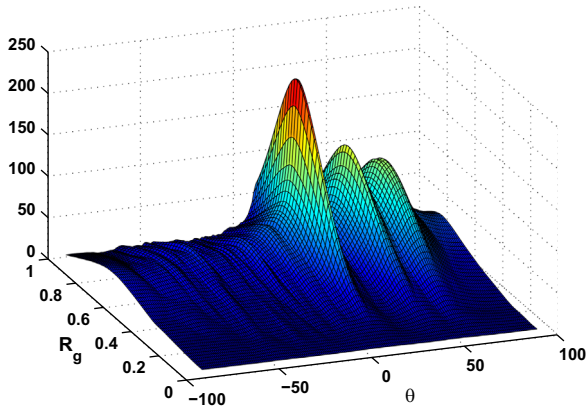


Fig. 4. 2-D spectrum for reverberant mixture of scenario 4.

presumed with zero-mean Gaussian distribution for both channels independent from each other. Hence, the likelihood of the parameters $\mathbf{Z} = \{\mathbf{W}, \mathbf{H}, \Phi\}$ is obtained as below:

$$p(\mathbf{X}_{ft}|\mathbf{Z}) = \prod_{i,f,t} \frac{1}{\pi\sigma^2} \exp\left(-\frac{|X_{ift} - Y_{ift}|}{\sigma^2}\right) \quad (13)$$

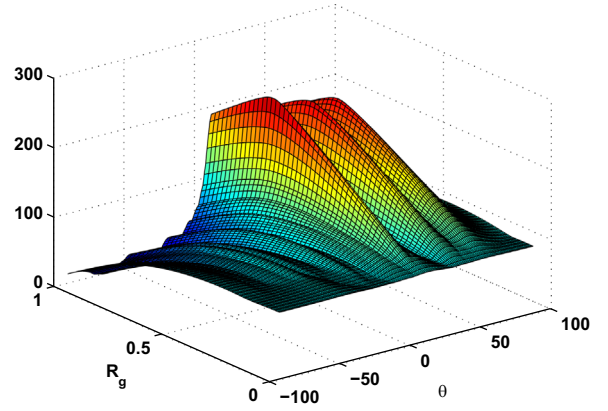


Fig. 5. 2-D spectrum for anechoic mixture of scenario 1 obtained without applying non-linearity.

Table 2

Percentage of correct estimation of the number of sources.

Number of sources	2	3	4	5	6	7	8	9	10
2-D spectrum-based	100	100	100	100	80	70	50	30	30
DEMIX-Anechoic	100	100	100	100	70	60	60	20	0

An iterative algorithm for maximizing the likelihood function is now developed. The derivation steps can be found in [Appendix A](#). The update relations are obtained as follows:

$$W_{fk} = \frac{\sum_{i,t} \frac{H_{kt}}{\beta_{kft}} \text{Re}[\bar{X}_{ikft}^* e^{i\phi_{kft}} A_{est,i}^{p_k}]}{\sum_{i,t} \frac{H_{kt}^2 |A_{est,i}^{p_k}|^2}{\beta_{kft}}}, \quad W_{fk} \leftarrow \frac{W_{fk}}{\sum_f W_{fk}} \quad (14)$$

$$H_{kt} = \frac{\sum_{i,f} \frac{W_{fk}}{\beta_{kft}} \text{Re}[\bar{X}_{ikft}^* e^{i\phi_{kft}} A_{est,i}^{p_k}]}{\sum_{i,f} \frac{W_{fk}^2 |A_{est,i}^{p_k}|^2}{\beta_{kft}}} \quad (15)$$

$$\sin(\phi_{kft}) = \frac{Q_{1kft}}{\sqrt{Q_{1kft}^2 + Q_{2kft}^2}} \quad (16)$$

$$\cos(\phi_{kft}) = \frac{Q_{2kft}}{\sqrt{Q_{1kft}^2 + Q_{2kft}^2}}$$

where the parameters β_{kft} , \bar{X}_{ikft} , Q_{1kft} and Q_{2kft} are defined in [Appendix A](#).

The β_{kft} parameter is set to $\beta_{kft} = W_{fk} H_{kt} / \sum_k W_{fk} H_{kt}$ at each iteration.

To avoid scaling ambiguity, the columns of \mathbf{W} are normalized. The iterative algorithm steps are executed in the following order:

- (1) The initial values of the set of parameters \mathbf{Z} are set to those obtained via factorizing the primary source estimates \mathbf{Z}^{bm} .
- (2) $\bar{\mathbf{X}}$ is updated according to (A.5) in the appendix.
- (3) The parameters $\mathbf{Z} = \{\mathbf{W}, \mathbf{H}, \Phi\}$ are updated according to Eqs. (14)–(16).

Table 3

Mixing parameters used to simulate mixtures with up to 10 sources.

Parameter	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
κ_j	0.62	1.09	0.87	1.74	1.35	1.23	0.95	1.53	0.72	0.54
δ_j	−0.41	−0.01	0.62	0.01	−0.16	0.56	−0.87	−1.1	0.32	0.15

(4) β is updated.

The complex spectrogram of the individual sources is ultimately reconstructed according to the following expression:

$$S_{jft} = \sum_{k \in K_j} W_{jk} H_{kt} e^{i\phi_{jft}} \quad (17)$$

The time-domain source signals can easily be derived by applying inverse STFT operation to (17).

4. Experiments

The mixture signal is synthetically generated using the Roomsim Toolbox [7] for a rectangular room of dimensions $6.25 \times 3.75 \times 2.5$ m and omnidirectional microphones. The source and microphone height are set to 1.1 m. The position of the microphone centers in (x, y) coordinates lies at (1.56 m, 1.87 m). We evaluate our proposed method under 4 different experimental settings. In the first and second scenarios, we consider anechoic environments. The respective positions of the microphones and sources in the room for scenarios 1, 3 and 4 are the same and shown in Fig. 1a. A different arrangement is considered in scenario 2 (Fig. 1b) accounting for the cases where the source signal may encounter a different attenuation in each channel. For scenarios 3 and 4, a moderate reverberation condition is simulated and the impulse responses are derived by assuming the reverberation times of $T_{60} = 50$ ms and $T_{60} = 100$ ms respectively using Campbell et al. [7]. The relation between room reverberation time and absorption coefficient of the surfaces given in Gustafsson et al. [12] is utilized for generating the impulse responses corresponding to each case.

The male and female speech signals are taken from dev2 dataset of the SiSEC'08 “underdetermined speech and music mixtures” task [28]. The common parameter setting for all scenarios is listed in Table 1.

4.1. Evaluating source counting and channel estimation

The 2-D spectra for a single synthetic mixture of 3 male speech sources are represented in Figs. 2–4, corresponding to scenarios 1, 2 and 4 respectively. The inferred peak locations for anechoic scenarios 1 and 2 are obtained as $R_{g1} = [0.72, 0.72, 0.72]$, $\theta_{g1} = [20^\circ, 40^\circ, 59^\circ]$ and $R_{g2} = [0.56, 0.8, 0.82]$, $\theta_{g2} = [-83^\circ, 47^\circ, 79^\circ]$ which shows a near perfect matching with the actual simulated impulse responses of the channel and source directions. The true R_g values for anechoic mixtures of scenarios 1 and 2 are $[0.719 \ 0.729 \ 0.737]$ and $[0.553 \ 0.819 \ 0.833]$ respectively. The need for considering a dimension corresponding to the amplitude response (R_g) in the spectrum is revealed when we deal

with cases similar to scenario 2 where the microphone spacing is larger and leads to different attenuation of the source signals received at each microphone.

Adding reverberation reduces the sparsity in the TF representation hence the peaks for scenario 4 are less sharp but they are still easily detectable by our peak finding algorithm. The corresponding channel parameters are obtained as $R_{g4} = [0.68, 0.72, 0.72]$, $\theta_{g4} = [21^\circ, 40^\circ, 59^\circ]$.

To measure the effectiveness of the nonlinear function applied in Eq. (8), we evaluate the 2-D spectrum for anechoic mixture of scenario 1 without applying the non-linearity. As can be seen in Fig. 5, the peaks are wider w.r.t. the 2-D spectrum in Fig. 2. Obviously, sharper peaks emerge in the spectrum when non-linearity is applied.

In order to evaluate our proposed source counting method, the rate of success of the algorithm in the estimation of the true number of sources is calculated and reported in Table 2. For this purpose, J sources are synthetically mixed together. This is done for J from 2 to 10. Taking the first channel as the reference, the mixing coefficient of the second channel corresponding to source j can be written as

$$a_{2j} = \kappa_j \exp\left(i2\pi f_s \frac{\delta_j}{f_s}\right) \quad (18)$$

where f_s is the sampling frequency, κ_j represents the relative mixing gain and δ_j denotes the relative time delay in samples. The mixing parameters are generated assuming $d_m = 1$ cm with different relative positions of the sources and microphones and are listed in Table 3. For each J , we generate 10 different mixtures by randomly selecting the original signals from speech/music sources of the dev2 dataset. The mixing parameters are obtained assuming an arrangement similar to scenario 2 with the difference that the source directions are assumed to be uniformly spaced in the interval $[0 \ \pi]$. The results are compared with the results of the DEMIX-Anechoic method [3] which are obtained using the software provided in Arberet [2]. Like in their work, the success rate is calculated as the percentage of correct estimations out of 10 trials. As can be observed, the algorithms perform perfectly up to 5 sources and the performance of our source counting algorithm is quite comparable and sometimes better than the DEMIX-Anechoic algorithm.

To also evaluate the mixing parameter estimation performance, the mean mixing parameter error (MMPE) is calculated similar to the mean direction error (MDE) proposed in Arberet et al. [3]. Given the true mixing parameter values $\kappa = [\kappa_1 \dots \kappa_J]$, $\delta = [\delta_1 \dots \delta_J]$ and the estimated ones $\hat{\kappa} = [\hat{\kappa}_1 \dots \hat{\kappa}_J]$, $\hat{\delta} = [\hat{\delta}_1 \dots \hat{\delta}_J]$ the MMPE is defined for both of the amplitude and delay parameters as below:

$$\text{MMPE}(\kappa, \hat{\kappa}) = \min_{P \in S_J} \frac{1}{J} \sum_{j=1}^J |\kappa_j - \hat{\kappa}_{P(j)}|$$

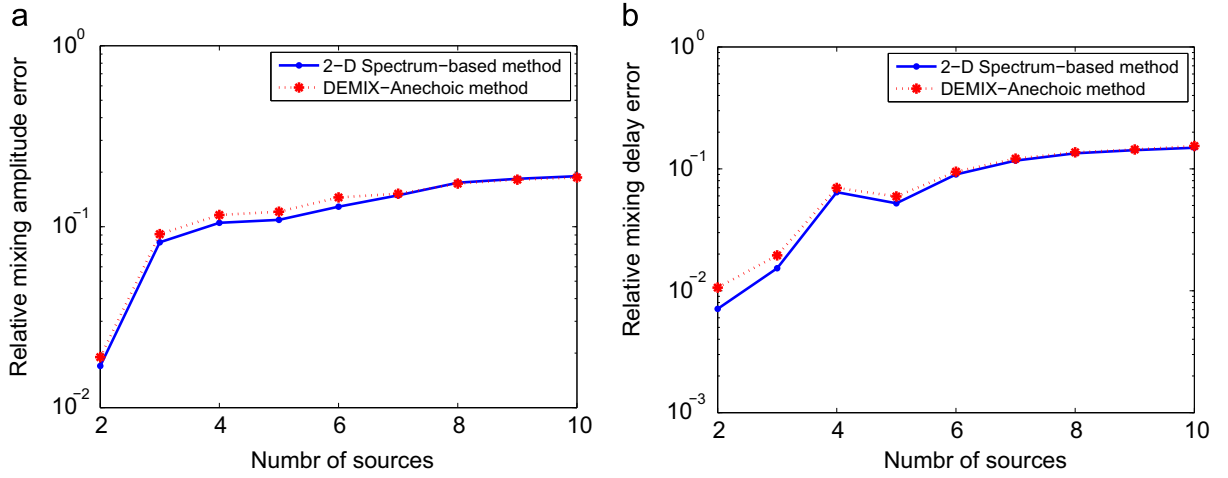


Fig. 6. Relative mean mixing parameter error (RMMPE) as a function of the number of sources: (a) Relative mixing amplitude error, (b) Relative mixing delay error.

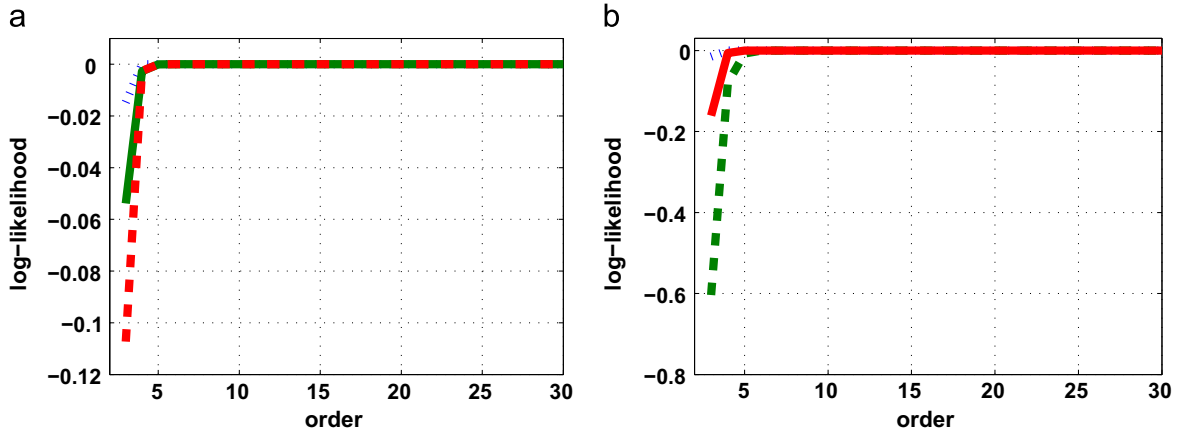


Fig. 7. Log-likelihood function against order for (a) male and (b) female sources.

Table 4
BSS Evaluation metrics obtained for scenario 1.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	12.8	18	14.3
Binary masking	8.9	17.7	10.1
l_0 -norm minimization	6.1	10.7	9.5

Table 5
BSS Evaluation metrics obtained for scenario 2.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	11.7	16.6	13.2
Binary masking	9	15.1	9.6
l_0 -norm minimization	5.2	8.4	9.5

Table 6
BSS Evaluation metrics obtained for scenario 3.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	9.7	11.6	13.8
Binary masking	7.5	12.2	9.3
l_0 -norm minimization	7.7	10	10.2

Table 7
BSS Evaluation metrics obtained for scenario 4.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	6.8	10.6	13.2
Binary masking	4.7	8.8	8.5
l_0 -norm minimization	4.9	6.7	11.8

$$\text{MMPE}(\delta, \hat{\delta}) = \min_{P \in S_J} \frac{1}{J} \sum_{j=1}^J |\delta_j - \hat{\delta}_{P(j)}| \quad (19)$$

where S_J is the permutation group of size J . Toward this goal, the J highest peaks of the 2-D spectrum found by the

peak finder algorithm are considered as the estimated mixing coefficient values of the sources. Similarly, for DEMIX-Anechoic algorithm, we set and fix the number of sources to J . To measure the error in terms of relative precision, the relative mean mixing parameter errors

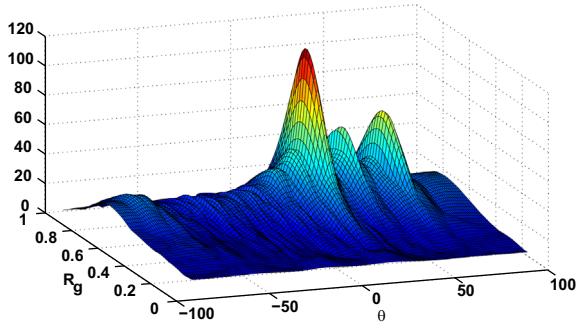


Fig. 8. 2-D spectrum for reverberant mixture with $T_{60} = 500$ ms.

(RMMPE) are defined as

$$\begin{aligned} \text{RMMPE}(\kappa, \hat{\kappa}) &= \frac{\text{MDE}(\kappa, \hat{\kappa})}{\min_{j \neq j'} |\kappa_j - \kappa_{j'}|} \\ \text{RMMPE}(\delta, \hat{\delta}) &= \frac{\text{MDE}(\delta, \hat{\delta})}{\min_{j \neq j'} |\delta_j - \delta_{j'}|} \end{aligned} \quad (20)$$

where the denominators denote the minimum distance between the true mixing parameter values. RMMPE values for amplitude and delay parameters are plotted in Fig. 6a and b respectively against the number of the sources. Because of the higher resolution needed for these experiments, we have used 18,000 θ segments and 100 R_g segments for evaluating the 2-D spectrum to achieve an angular resolution of 0.01° and normalized amplitude resolution of 0.01. It can be seen that our proposed algorithm outperforms DEMIX-Anechoic method up to 6 sources in terms of delay parameter relative error and up to 7 sources in terms of amplitude parameter relative error. Adding more sources, the two algorithms show nearly the same performance.

4.2. Model order selection

To get an insight of the proper number of components needed for modeling each source, a model order selection scheme is implemented here. This method is applied to the primary source spectrograms given by binary masking. The initial values of the elements of \mathbf{W}^{bm} and \mathbf{H}^{bm} are drawn randomly from the absolute value of a standard Gaussian distribution plus 1 ($\text{abs}(N(0, 1)) + 1$). The \mathbf{W}^{bm} elements are normalized according to 14. Initial values of Φ^{bm} are drawn from a uniform distribution over the interval $[-\pi, \pi]$. The update algorithm is iterated 500 times. The log-likelihood function is evaluated against a range of order values $k \in \{3 \dots 30\}$. The likelihood graphs for the mixture of 3 male and 3 female speech signals generated under scenario 1 are represented in Fig. 7a and b respectively. We exploit the BIC metric defined as

$$\text{BIC} = -2\text{LL} + N_p \log(N_{\text{obs}}) \quad (21)$$

where LL represents the corresponding log-likelihood vector, N_p denotes the number of model parameters and N_{obs} is the number of observed samples associated with

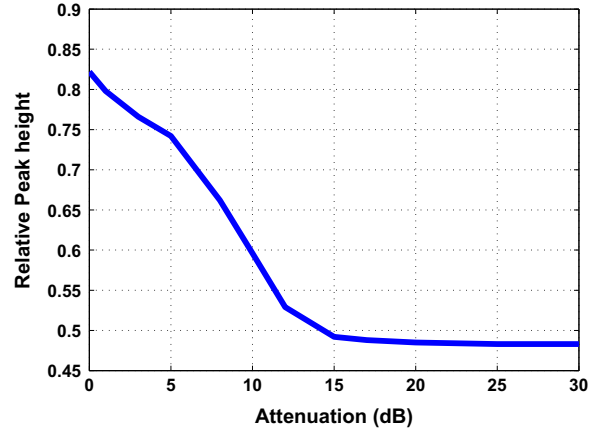


Fig. 9. Relative peak height against the source signal attenuation for reverberant mixture with $T_{60} = 500$ ms.

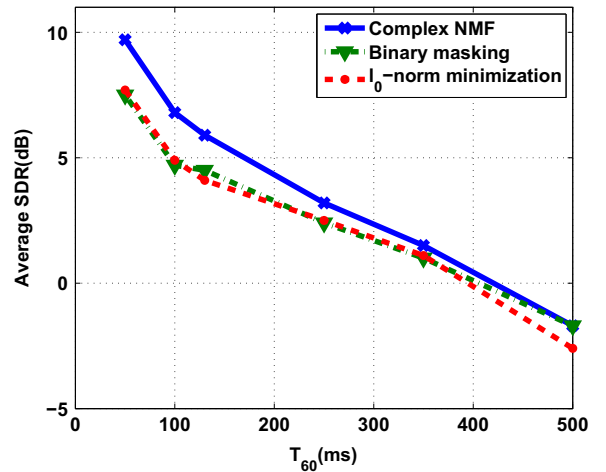


Fig. 10. Average SDR as a function of the reverberation time.

each value in LL. The order value corresponding to the minimum obtained BIC metric is taken as the optimal number of model components. This leads to order values of [5 6 6] and [7 8 7] for the male and female sources, respectively. For mitigating the computational burden, we do not re-execute the model order selection algorithm for each mixture in the next step. Instead, we take a fixed number of components considered for each source in our model based on the above experiments ($|K_j| = 8, j = 1 \dots J$).

4.3. Source separation evaluation

Here, we are to evaluate the performance of our proposed source separation algorithm using the extended multichannel model. We generated 10 mixtures of 3 (male and/or female) speech signals using the impulse responses obtained through Campbell et al. [7]. The separation quality is measured by calculating the evaluation metrics including Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR) and Signal to Artifact Ratio (SAR) [29].

Table 8Average performance metrics obtained for $T_{60} = 150$ ms.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	8.5	12.2	15.0
Full-rank method with our proposed initialization scheme	6.9	10.8	10.5
Full-rank method with hierarchical clustering-based initialization	6.2	10.0	9.4

Table 9Average performance metrics obtained for $T_{60} = 500$ ms.

Method	SDR (dB)	SIR (dB)	SAR (dB)
Complex NMF	−1.9	0.3	5.2
Full-rank method with our proposed initialization scheme	4.2	5.3	8.9
Full-rank method with hierarchical clustering-based initialization	3.61	4.9	8.1

The performance measures averaged over all sources and all mixtures are reported in Tables 4–7 corresponding to 4 scenarios. As comparison materials, the results of binary masking and l_0 -norm minimization algorithms are also evaluated using the reference software in Vincent et al. [28] and are listed in the tables. It can be observed that our proposed method outperforms both of the state-of-the-art methods in all evaluation metrics. The algorithm works also in moderate reverberant environments since the sparsity in the TF domain is still sufficient for pure peaks to emerge in the 2-D spectrum.

5. Discussion

We now investigate to what extent the algorithm performance is still acceptable when we are to handle longer than the intended reverberation times. Toward this objective, the effectiveness is first measured in the first stage, i.e. the channel estimation and source counting, and then source separation performance is assessed against different T_{60} values.

Fig. 8 represents the 2-D spectrum for the arrangement of Fig. 1a with $T_{60} = 500$ ms. The proposed spectrum-based approach is apparently still applicable in highly reverberant environments for the purpose of source counting and angle of arrival estimation. The corresponding channel parameters are obtained as $R_g = [0.7, 0.7, 0.66]$, $\theta_g = [20^\circ, 40^\circ, 59^\circ]$ in this case. As can be observed, the peaks of the spectrum are getting wider by increasing the reverberation time but the number of the sources and source directions are obtained with the same accuracy as the moderately reverberant case according to our experimentation. To also investigate the effectiveness of the relative amplitude threshold (0.5) in this highly reverberated case, we reduce the strength of one source in the mixture of 3 sources and plot the relative peak height corresponding to that source as a function of the source strength (the attenuation applied to the signal). The result is shown in Fig. 9. It can be observed that the chosen fixed threshold value (0.5) is quite effective in detecting the peak up to 12 dB reduction in the source signal amplitude. From this analysis, we can state that the parameter choice will be related to the signal conditions (reverberation and SNR) in a practical deployment.

The performance of the source separation is shown in Fig. 10 in terms of average SDR. The average performance metrics are calculated by taking the average over all sources and all mixtures in the dev2 database (speech/music). The performance degrades when increasing the reverberation time for all of the methods. This is an expected result since there is no model for multipath propagation, but complex NMF still outperforms the other methods. Nevertheless, the satisfying results of the first stage can be exploited in a suitable framework for source separation in highly reverberant conditions. Taking reverberation into account, most of the source separation methods work based on the spatial covariance matrix modeling [4,9] and use the Expectation Maximization (EM) algorithm for maximizing the data likelihood. The EM algorithm is very sensitive to initialization. The mixing coefficients obtained in the first step could provide proper initialization for these approaches. For this purpose, we propose to use the estimated channel parameters given in (10). We have obtained the average performance metrics for the full-rank method of Duong et al. [9]. The metrics are evaluated for two different reverberation conditions ($T_{60} = 150$ ms and $T_{60} = 500$ ms) represented in Tables 8 and 9 respectively. We also compare our proposed initialization scheme with the hierarchical clustering-based initialization scheme proposed in [9]. We execute 50 EM algorithm iterations and the number of clusters for hierarchical clustering is set to 30. As it can be observed, in low reverberation condition ($T_{60} = 150$ ms), our proposed method outperforms that of Duong et al. [9] with both initialization schemes. It is also evident that our proposed initialization makes the EM algorithm to lead to better results. In higher reverberation, the full-rank method is performing better than the complex NMF approach and again better performance is achieved using our proposed initialization scheme.

6. Conclusion

In the first step, we developed a novel scheme for source counting and channel estimation in anechoic or moderate reverberant environments introducing a 2-D spectrum over both phase and attenuation difference between the channels.

Secondly, an extension of the complex factorization framework was developed for the purpose of source separation which uses the channel mixing coefficients estimated in previous step. The initial values of the factors are set to those obtained from primary source spectrogram estimates given by binary masking. The average BSS performance metrics shows the superiority of our algorithm over all three evaluation criteria for all different considered source separation scenarios.

We have also introduced a model order selection scheme using the BIC metric for inferring the required number of components for modeling each source. This algorithm can be applied to the primary source spectrogram estimates.

Assuming a prior distribution for the \mathbf{H} parameters can be helpful if we set up a training procedure for obtaining the optimal values of the hyperparameters of the prior distribution and can be considered in future work for the purpose of performance improvement.

Acknowledgments

This research was funded by the KU Leuven research Grant GOA/14/005 (CAMETRON).

Appendix A. Derivation of the update relations for the extended multichannel model

We are to solve the following optimization problem:

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & f(\mathbf{Z}) = \sum_{i,f,t} |X_{ift} - Y_{ift}|^2 \\ \text{subject to} \quad & \sum_f W_{fk} = 1 \quad (k = 1 \dots K) \end{aligned} \quad (\text{A.1})$$

We define the following auxiliary function with auxiliary parameters $\bar{\mathbf{Z}} = \{\bar{\mathbf{X}}\}$ inspired from Kameoka et al. [15]

$$f^+(\mathbf{Z}, \bar{\mathbf{Z}}) = \sum_{i,k,f,t} \frac{|\bar{X}_{ikft} - W_{fk} H_{kt} e^{j\phi_{kft}} A_{\text{est},i}^{p_k}|^2}{\beta_{kft}} \quad (\text{A.2})$$

where β_{kft} can be any positive number satisfying $\sum_k \beta_{kft} = 1$. The auxiliary function $f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ should satisfy

$$f(\mathbf{Z}) = \min_{\bar{\mathbf{Z}}} f^+(\mathbf{Z}, \bar{\mathbf{Z}}) \quad (\text{A.3})$$

If the above condition is satisfied, it can be shown based on Lee and Seung [16] that $f(\mathbf{Z})$ is non-increasing under the updates $\bar{\mathbf{Z}} \leftarrow \arg\min_{\bar{\mathbf{Z}}} f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ and $\mathbf{Z} \leftarrow \arg\min_{\mathbf{Z}} f^+(\mathbf{Z}, \bar{\mathbf{Z}})$.

$f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ is an auxiliary function for $f(\mathbf{Z})$ if

$$\sum_k \bar{X}_{ikft} = X_{ift} \quad (\text{A.4})$$

$f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ is minimized w.r.t. $\bar{\mathbf{Z}}$ when

$$\bar{X}_{ikft} = W_{fk} H_{kt} e^{j\phi_{kft}} A_{\text{est},i}^{p_k} + \beta_{kft} (X_{ift} - Y_{ift}) \quad (\text{A.5})$$

Eq. (A.5) can be derived by adding the Lagrange multiplier term to the auxiliary function of Eq. (A.2):

$$\Lambda(\bar{\mathbf{X}}, \lambda) = f^+(\mathbf{Z}, \bar{\mathbf{Z}}) + \lambda_{ift} \left(\sum_k \bar{X}_{ikft} - X_{ift} \right) \quad (\text{A.6})$$

and then taking the derivative of (A.6) w.r.t. \bar{X}_{ikft} and setting it to zero, we have

$$2\bar{X}_{ikft} - 2W_{fk} H_{kt} e^{j\phi_{kft}} A_{\text{est},i}^{p_k} = -\lambda_{ift} \beta_{kft} \quad (\text{A.7})$$

λ_{ift} is obtained by summation over k on both sides of (A.7) and applying the constraint (A.4):

$$\lambda_{ift} = 2(X_{ift} - Y_{ift}) \quad (\text{A.8})$$

Consequently, \bar{X}_{ikft} is obtained as given by (A.5). Substituting (A.5) in the expression (A.2), we determine the minimum value of the auxiliary function which is equal to $f(\mathbf{Z})$, so condition (A.3) is satisfied.

Differentiating $f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ w.r.t. W_{fk} and H_{kt} and setting them to zero, leads to the update relations for these parameters as stated in (14) and (15).

Deriving the update relations for ϕ_{kft} is not as straightforward as the above parameters. The first derivative of the auxiliary function $f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ w.r.t. ϕ_{kft} is proportional to the following expression:

$$\begin{aligned} \frac{\partial f^+(\mathbf{Z}, \bar{\mathbf{Z}})}{\partial \phi_{kft}} &\propto \sum_i \left\{ W_{fk} H_{kt} \text{Re}(A_{\text{est},i}^{p_k} \bar{X}_{ikft}^*) \right\} \sin(\phi_{kft}) \\ &+ \left\{ W_{fk} H_{kt} \text{Im}(A_{\text{est},i}^{p_k} \bar{X}_{ikft}^*) \right\} \cos(\phi_{kft}) \end{aligned} \quad (\text{A.9})$$

Since we assumed the first channel as reference, the channel coefficient on the first channel, $A_{\text{est},1}^{p_k} = 1$ in accordance with (10). Thus, setting (A.9) to zero, we obtain the following update relation for ϕ_{kft} coefficients:

$$\begin{aligned} Q_{1kft} &= \text{Im}(\bar{X}_{1kft}) - \text{Im}(A_{\text{est},2}^{p_k} \bar{X}_{2kft}^*) \\ Q_{2kft} &= \text{Re}(\bar{X}_{1kft}) + \text{Re}(A_{\text{est},2}^{p_k} \bar{X}_{2kft}^*) \\ \phi_{kft} &= \text{atan}\left(\frac{Q_{1kft}}{Q_{2kft}}\right) \end{aligned} \quad (\text{A.10})$$

Actually, the above solution does not result in a unique value for ϕ_{kft} . The unique solution is achieved by noticing the point that the second derivative of the cost function $f^+(\mathbf{Z}, \bar{\mathbf{Z}})$ w.r.t. ϕ_{kft} should be positive:

$$\begin{aligned} \frac{\partial^2 f^+(\mathbf{Z}, \bar{\mathbf{Z}})}{\partial \phi_{kft}^2} &\propto \sin(\phi_{kft}) \cdot \left\{ \text{Im}(\bar{X}_{1kft}) - \text{Im}(A_{\text{est},2}^{p_k} \bar{X}_{2kft}^*) \right\} \\ &+ \cos(\phi_{kft}) \cdot \left\{ \text{Re}(\bar{X}_{1kft}) + \text{Re}(A_{\text{est},2}^{p_k} \bar{X}_{2kft}^*) \right\} \geq 0 \end{aligned} \quad (\text{A.11})$$

Satisfying the above condition will specify the unique solution inferred from (A.10). It can be implied from (A.11) that the sine and cosine of ϕ_{kft} should have the same signs as the nominator and denominator of (A.10) respectively. The form of the update relation stated in (16) will satisfy this.

References

- [1] S. Araki, H. Sawada, R. Mukai, S. Makino, A novel blind source separation method with observation vector clustering, in: Proceedings of IWAENC, 2005, pp. 117–120.
- [2] S. Arberet, Demix-anechoic Software Reference Website. URL <https://sites.google.com/site/simonarberet/codes/>.
- [3] S. Arberet, R. Gribonval, F. Bimbot, A robust method to count and locate audio sources in a multichannel underdetermined mixture, IEEE Trans. Signal Process. 58 (1) (2010) 121–133.
- [4] S. Arberet, A. Ozerov, N.Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, P. Vanderghenst, Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation, in: 2010 10th International Conference on Information

- Sciences Signal Processing and their Applications (ISSPA), IEEE, Kuala Lumpur, Malaysia, 2010, pp. 1–4.
- [5] N. Bertin, R. Badeau, E. Vincent, Fast Bayesian nmf algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription, in: WASPAA, 2009, pp. 29–32.
 - [6] P. Bofill, M. Zibulevsky, Underdetermined blind source separation using sparse representations, *Signal Process.* 81 (11) (2001) 2353–2362.
 - [7] D. Campbell, K. Palomaki, G. Brown, A matlab simulation of “shoe-box” room acoustics for use in research and teaching, *Comput. Inf. Syst.* 9 (3) (2005) 48.
 - [8] O. Dikmen, A.T. Cemgil, Unsupervised single-channel source separation using bayesian nmf, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA’09. IEEE, New Paltz, New York, USA, 2009, pp. 93–96.
 - [9] N.Q. Duong, E. Vincent, R. Gribonval, Under-determined reverberant audio source separation using a full-rank spatial covariance model, *IEEE Trans. Audio Speech Lang. Process.* 18 (7) (2010) 1830–1840.
 - [10] C. Févotte, A. Ozerov, Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues, *Explor. Music Contents* (2011) 102–115.
 - [11] D. FitzGerald, M. Cranitch, E. Coyle, Non-negative tensor factorisation for sound source separation, in: Proceedings of the Irish Signals and Systems Conference, Dublin Institute of Technology, Dublin, Ireland, 2005.
 - [12] T. Gustafsson, B.D. Rao, M. Trivedi, Source localization in reverberant environments: modeling and statistical analysis, *IEEE Trans. Speech Audio Process.* 11 (6) (2003) 791–803.
 - [13] M.Z. Ikram, D.R. Morgan, A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation, in: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, IEEE, Orlando, Florida, 2002, pp. 1–881.
 - [14] R. Jaiswal, D. FitzGerald, D. Barry, E. Coyle, S. Rickard, Clustering nmf basis functions using shifted nmf for monaural sound source separation, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 245–248.
 - [15] H. Kameoka, N. Ono, K. Kashino, S. Sagayama, Complex nmf: a new sparse representation for acoustic signals, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE, Taipei, Taiwan, 2009, pp. 3437–3440.
 - [16] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2001, pp. 556–562.
 - [17] B. Loesch, B. Yang, Blind source separation based on time–frequency sparseness in the presence of spatial aliasing, in: Latent Variable Analysis and Signal Separation, Springer, St. Malo, France, 2010, pp. 1–8.
 - [18] Y. Mitsufuji, A. Roebel, Sound source separation based on non-negative tensor factorization incorporating spatial cue as prior, New Paltz, New York, USA, knowledge, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 71–75.
 - [19] F. Nesta, M. Omologo, P. Svaizer, Multiple tdoa estimation by using a state coherence transform for solving the permutation problem in frequency-domain bss, in: IEEE Workshop on Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE, Cancun, Mexico, 2008, pp. 43–48.
 - [20] A. Ozerov, C. Févotte, Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation, *IEEE Trans. Audio Speech Lang. Process.* 18 (3) (2010) 550–563.
 - [21] V.G. Reju, S.N. Koh, I.Y. Soon, Underdetermined convolutive blind source separation via time–frequency masking, *IEEE Trans. Audio Speech Lang. Process.* 18 (1) (2010) 101–116.
 - [22] R. Saab, O. Yilmaz, M.J. McKeown, R. Abugharbieh, Underdetermined anechoic blind source separation via lq-basis-pursuit, *IEEE Trans. Signal Process.* 55 (8) (2007) 4004–4017.
 - [23] H. Sawada, S. Araki, R. Mukai, S. Makino, Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation, *IEEE Trans. Audio Speech Lang. Process.* 15 (5) (2007) 1592–1604.
 - [24] H. Sawada, H. Kameoka, S. Araki, N. Ueda, Multichannel extensions of non-negative matrix factorization with complex-valued data, *IEEE Trans. Audio Speech Lang. Process.* 21 (5) (2013) 971–982.
 - [25] H. Sawada, R. Mukai, S. Araki, S. Makino, A robust and precise method for solving the permutation problem of frequency-domain blind source separation, *IEEE Trans. Speech Audio Process.* 12 (5) (2004) 530–538.
 - [26] P. Smaragdis, J. Brown, Non-negative matrix factorization for polyphonic music transcription, in: 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003, pp. 177–180.
 - [27] E. Vincent, Complex nonconvex l_p norm minimization for underdetermined source separation, in: Independent Component Analysis and Signal Separation, Springer, London, UK, 2007, pp. 430–437.
 - [28] E. Vincent, S. Araki, P. Bofill, The 2008 signal separation evaluation campaign: a community-based approach to large-scale evaluation, in: Independent Component Analysis and Signal Separation, Springer, Paraty, Brazil, 2009, pp. 734–741.
 - [29] E. Vincent, R. Gribonval, C. Févotte, Performance measurement in blind audio source separation, *IEEE Trans. Audio Speech Lang. Process.* 14 (4) (2006) 1462–1469.
 - [30] T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, *IEEE Trans. Audio Speech Lang. Process.* 15 (3) (2007) 1066–1074.
 - [31] S. Winter, W. Kellermann, H. Sawada, S. Makino, Map-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l_1 -norm minimization, *EURASIP J. Appl. Signal Process.* 2007 (1) (2007) 81.
 - [32] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time–frequency masking, *IEEE Trans. Signal Process.* 52 (7) (2004) 1830–1847.